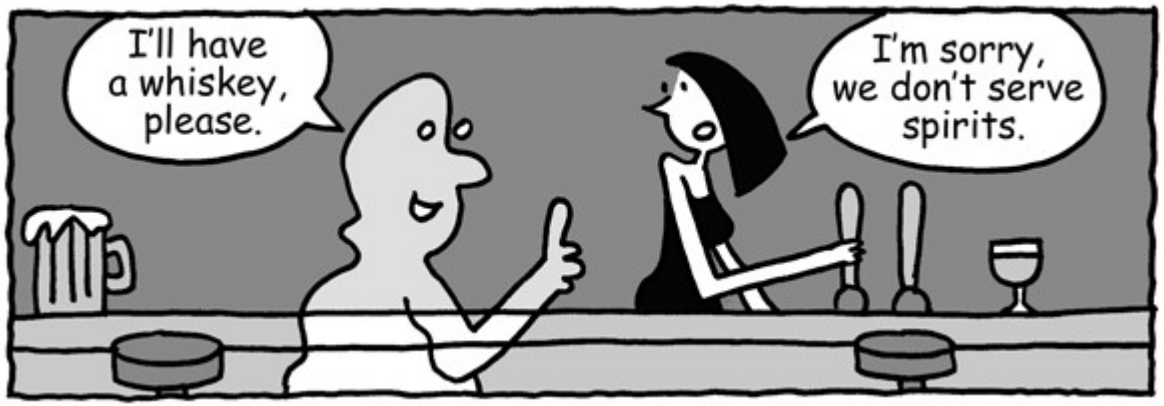
Word2Vec theory



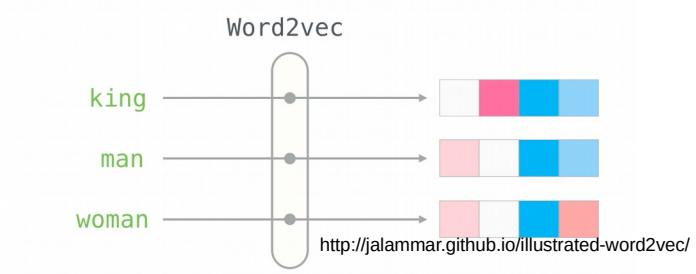
Word as combination of aspects

- Word are a representation of a concept
- Concept can have certain aspects.
- Characterize a concept behind a word:

	alive	**electrical**	**soft**	**big**	**...**	...	...
cat	yes	no	yes	yes	no	yes	no
table	No	no	no	no	no	yes	no
book	no	no	no	no	no	yes	yes
phone	no	yes	no	no	yes	no	yes

Word2vec: define aspects of a word

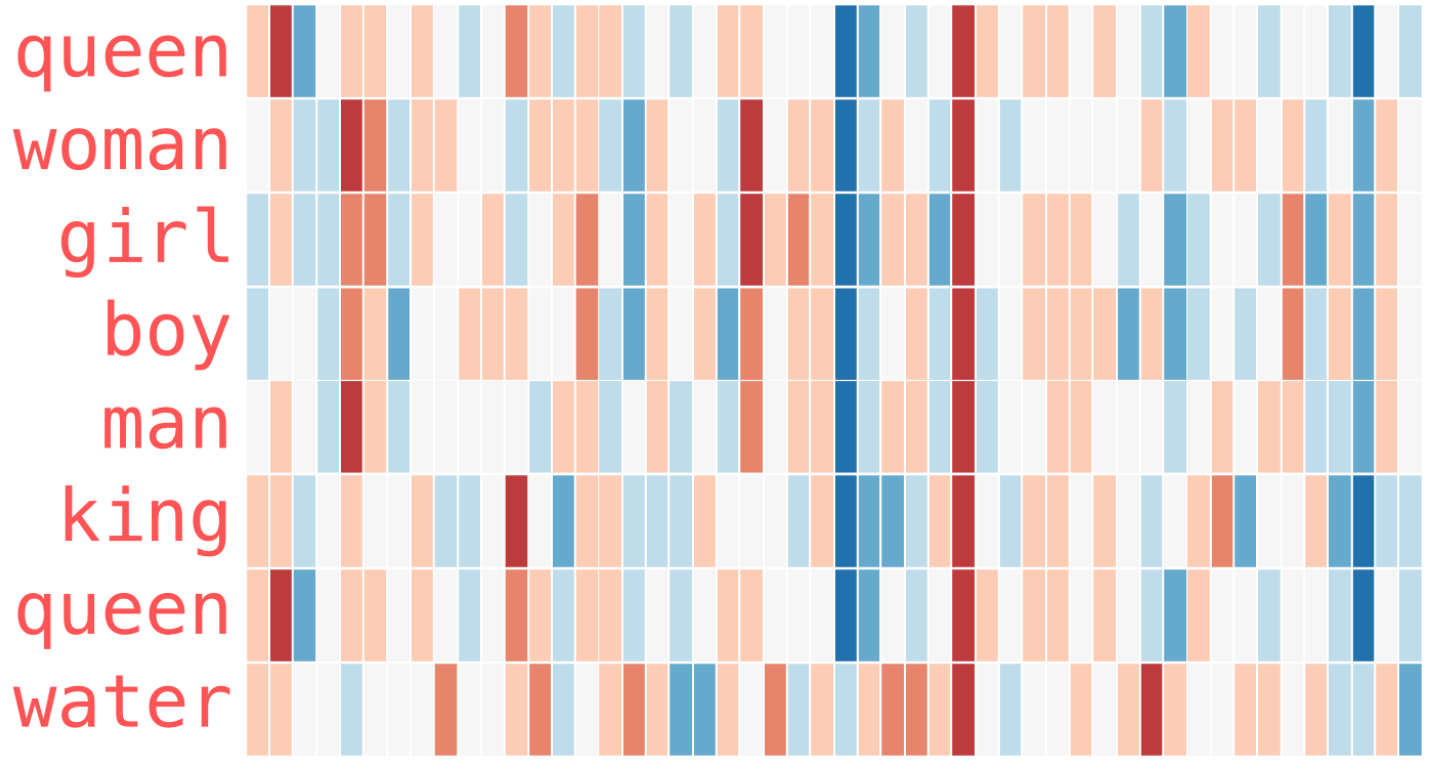
- Create a vector to capture the aspects
- Can have continuous (0 to 1) aspects instead of binary (yes/no)
- Aspects can be abstract



	alive	**electrical**	**soft**	**big**	***likable***	***abstract***	***value***
cat	0.95	0.2	0.8	0.3	0.9	0.1	0.8
table	0.0	0.1	0.2	0.4	0.2	0.1	0.1
book	0.01	0.05	0.3	0.1	0.8	0.2	0.2
phone	0.05	0.9	0.1	0.05	0.7	0.2	0.7
thought	0.01	0.01	0.01	0.01	0.1	0.8	0.9

Exercise: look at some trends

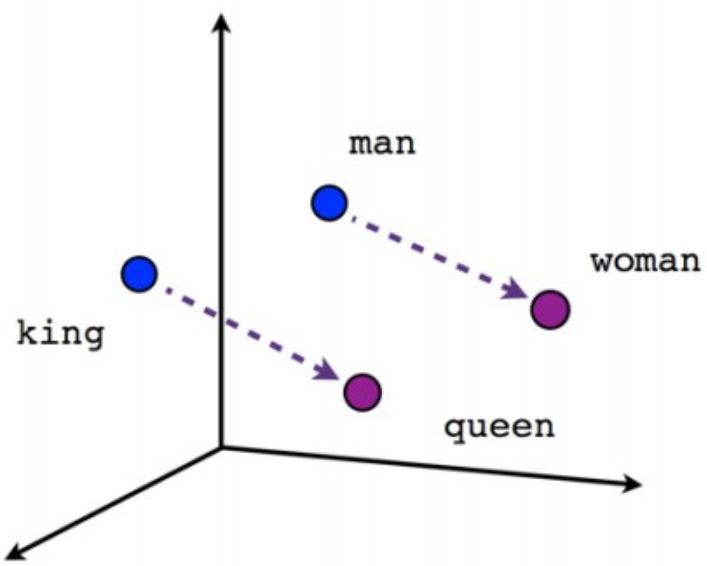
- Which feature is ‘human’ ?
- What separates man from boy / woman from girl?



http://jalammar.github.io/illustrated-word2vec/

Properties to vectors

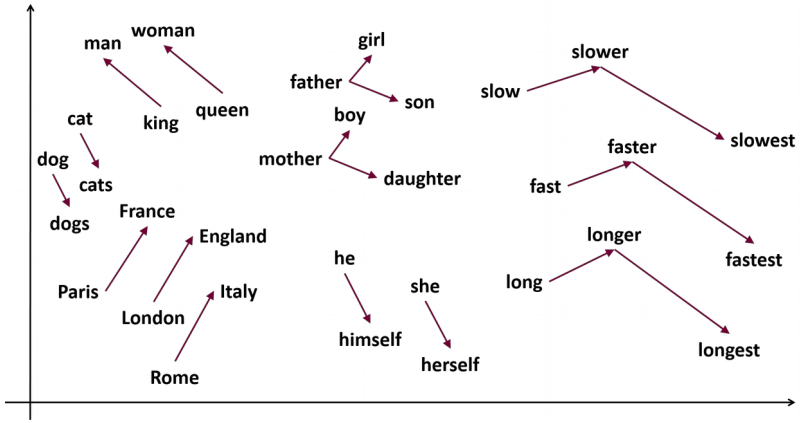
- Create an N-dimensional graph
- Can place each word on this N-dimensional graph



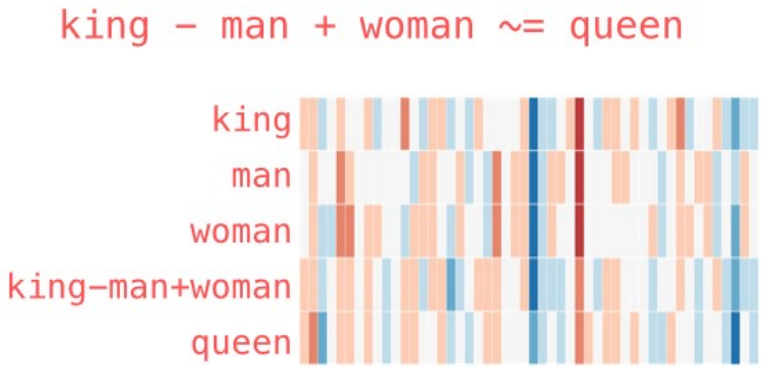
https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Relating properties / vectors

- Can create vectors between concepts
  - Mathematical operations possible (woman + vector_man → king = queen)
  - Similar concepts have similar features (cosine similarity)



https://samyzaf.com/ML/nlp/nlp.html



How to do this?

- In short: translate word to property space
- Question: How will we do this?
  - Need to think of features of a word
  - Need to fill in these features for each word there is
  - Sounds like a lot of work to do by hand...

Whats in a name?

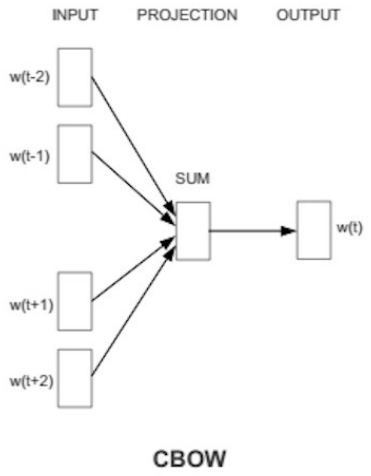
"You shall know a word by the company it keeps."

—J.R. Firth

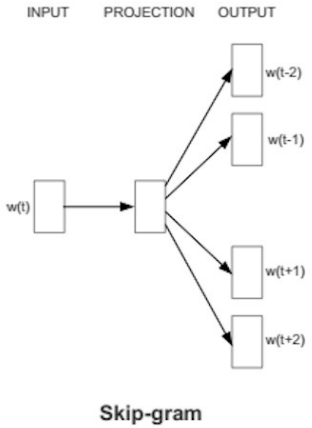
- Word defined by its context
  - A chair is something we sit on; A chair is something we work on
  - A cat is something we love; A cat is something we feed; A cat has personality
- Question: What is a **spirit**?
  - A spirit is see trough; a spirit is scary; a spirit haunts
  - A spirit is liquid; a spirit has alcohol; a spirit is nice
- The context of a word will tell you about the properties of a word

Methods to define words by context

- Continuous Bag of Words
  - Context to word
  - Good for small dataset
- Skip-gram
  - Word to context
  - Good for big dataset



CBOW



Skip-gram

https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

CBOW getting dataset

- Multiple in → one out
- Sliding window approach
-

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

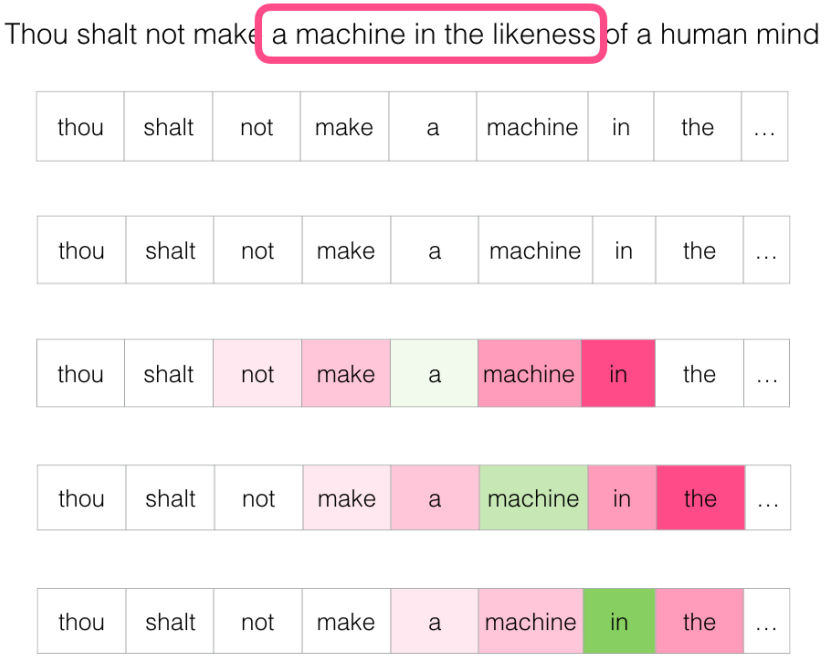
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make
not	make	a
make	a	machine
a	machine	in

skipgram getting dataset

- One in → multiple out
- Sliding window approach



input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

Add negative sampling

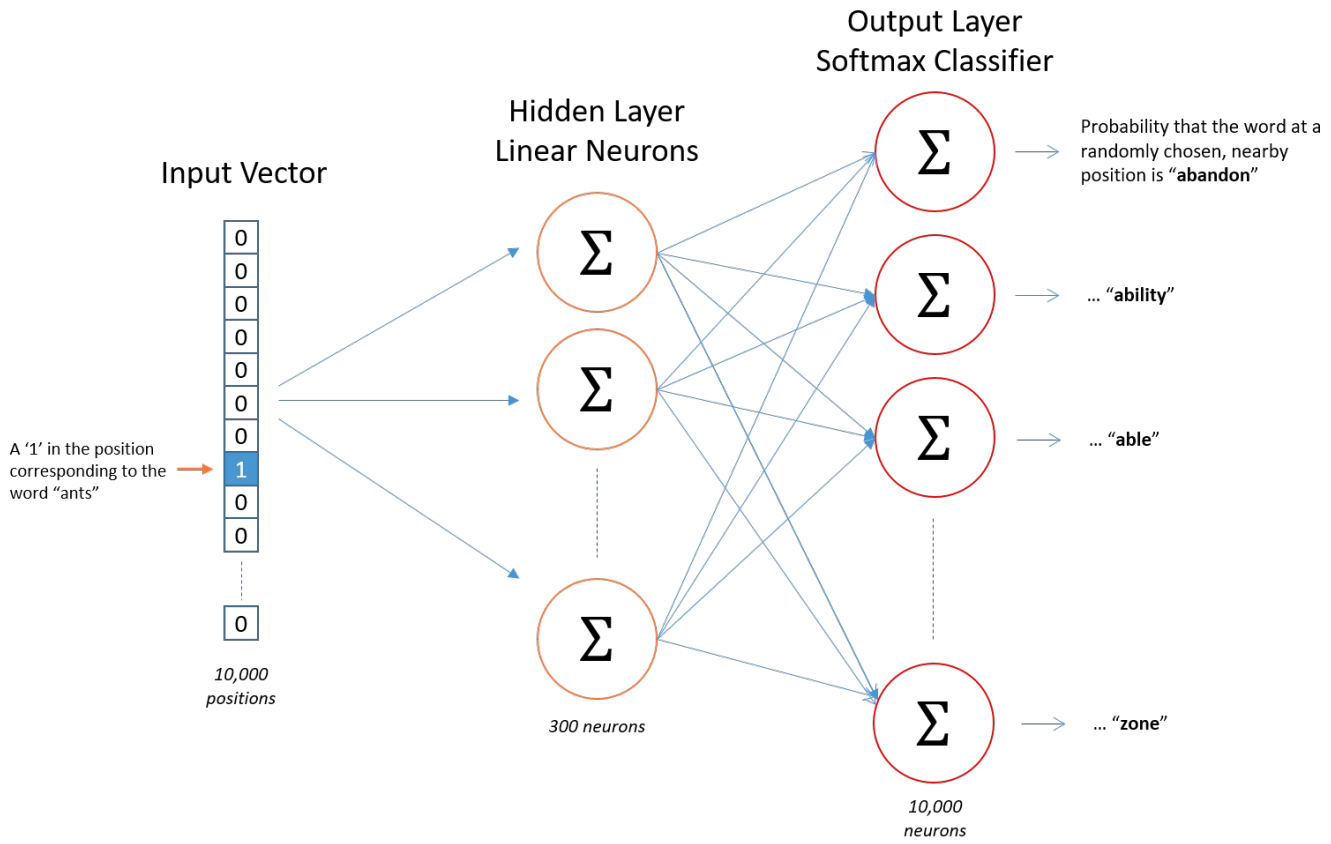
- 1: they are context
- 0: they are not context

dataset

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
...	...	...

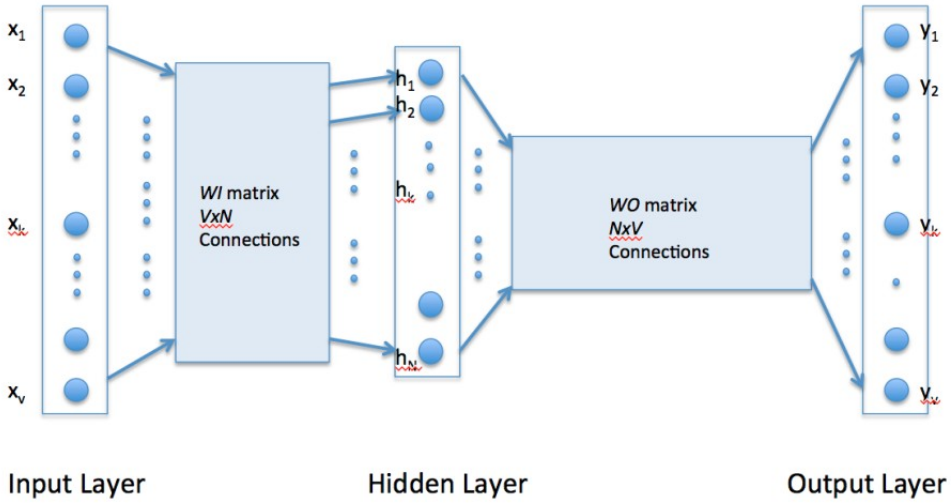
Architecture - skipgram

- Hidden layer serves as embedding
-



Architecture again

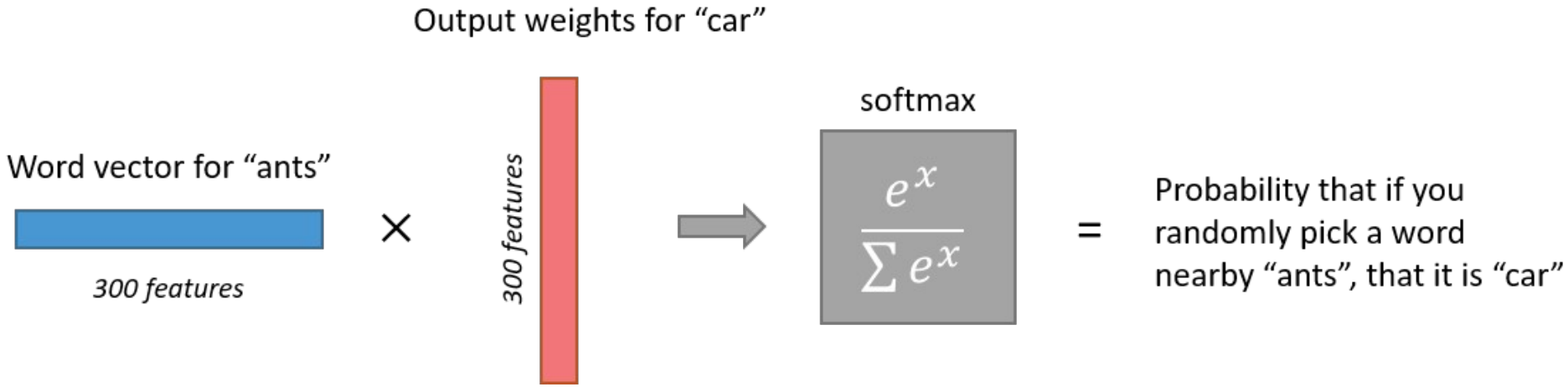
- First layer: input to embedding
- Second layer: embedding to output



The embedding / hidden layer

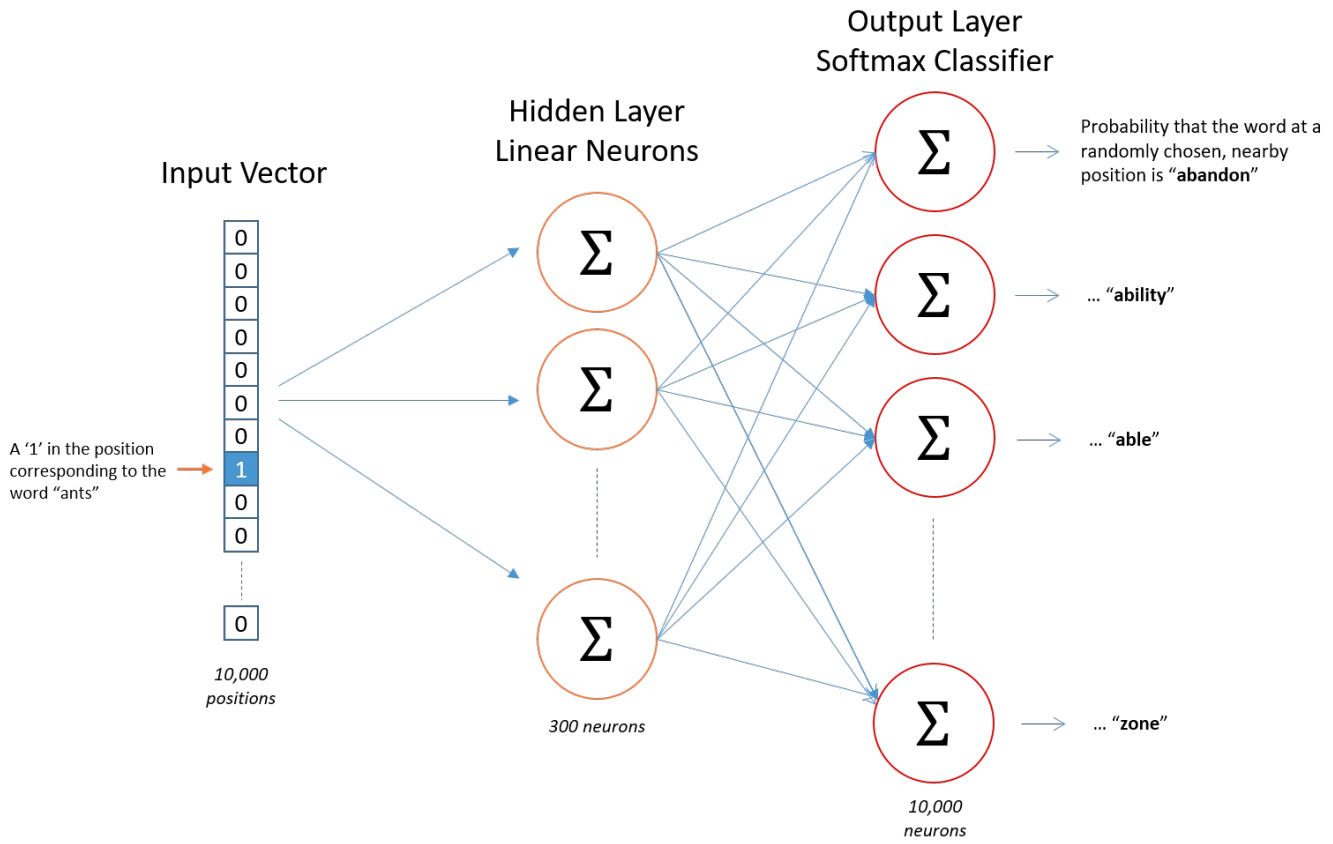
- One-hot X embedding = embedding

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$



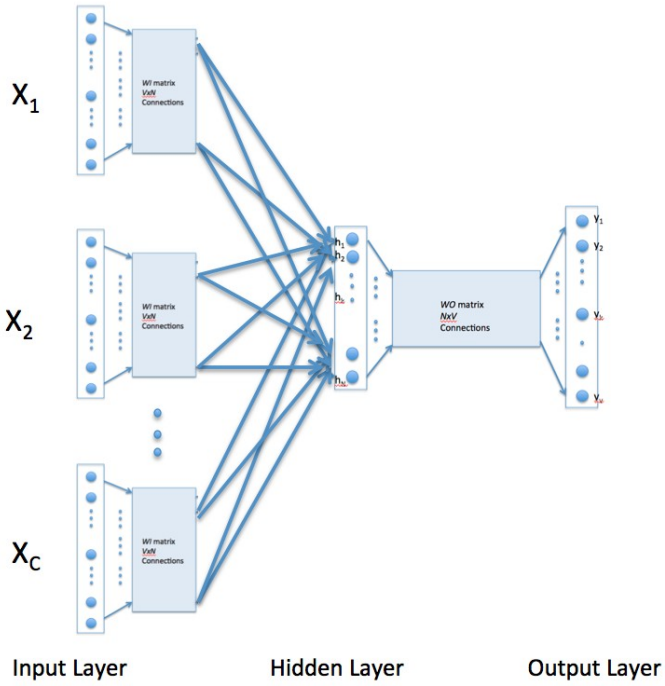
Training

- Minimize the liklihood of a negative sample
- Maximize the likelihood of a positive sample



How to do it for CBOW?

- Sum the input vectors, divide by the amount of input vectors
  - Get the context as an average word
- Train in the same way



End of presentation